Biostatistics (2017), **0**, 0, *pp.* 1–28
doi:10.1093/biostatistics/Paper16261

Penalized likelihood estimation of a trivariate additive probit model

PANAGIOTA FILIPPOU*

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

panagiota.filippou.12@ucl.ac.uk

GIAMPIERO MARRA

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

ROSALBA RADICE

Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

SUMMARY

This paper proposes a penalized likelihood method to estimate a trivariate probit model, which accounts for several types of covariate effects (such as linear, nonlinear, random and spatial effects), as well as error correlations. The proposed approach also addresses the difficulty in estimating accurately the correlation coefficients, which characterize the dependence of binary responses conditional on covariates. The parameters of the model are estimated within a penalized likelihood framework based on a carefully structured trust region algorithm with integrated automatic multiple smoothing parameter selection. The relevant numerical computation can be easily carried out using the `SemiParTRIV()` function in a freely available `R` package. The proposed method is illustrated through a case study whose aim is to model jointly adverse

*To whom correspondence should be addressed.

birth binary outcomes in North Carolina.

Key words: additive predictor, correlation-based penalty, penalized regression spline, simultaneous parameter estimation, trivariate probit model.

1. INTRODUCTION

Regression models usually involve one response variable and a set of covariates. However, modeling simultaneously more responses in a regression setting can be of considerable empirical relevance. The particular case of trivariate models has been discussed in the literature in various applied and methodological contexts (see, for instance, Genest et al., 2013; Król et al., 2016; Zhang et al., 2015; Zhong et al., 2012, and references therein).

This paper is about trivariate probit models which can be traced back to the seminal article by Ashford & Sowden (1970) on multivariate probit models. Chib & Greenberg (1998) later proposed a Bayesian approach for estimating such models. In these works, non-parametric covariates effects are not allowed for and the difficulty in estimating accurately the model's correlation coefficients at small or modest sample sizes is neither discussed nor dealt with. We address the first issue by considering trivariate probit models with additive or semi-parametric predictors, hence allowing for several types of covariate effects (such as linear, non-linear, random and spatial effects). This may help uncover interesting structures in the data and reduce the risk and consequences of mis-specifying covariate-response relationships (e.g., Donat & Marra, 2016, and references therein). The second issue is dealt with by introducing an approach for penalizing the correlation coefficients, which characterize the dependence of the binary responses conditional on regressors. Estimating such parameters accurately is crucial to obtain unbiased joint outcome probabilities, for instance. To implement these advances a reliable estimation algorithm needs to be developed. To this end, we extend to this context the penalized likelihood framework based on a trust region method with automatic smoothing parameter selection developed by Marra et al. (2016). Such extension relies

on the availability of the analytical score and Hessian components of the model's log-likelihood, which are derived in this paper and represent a contribution in itself. Asymptotic arguments of the proposed estimator are also provided. Note that in the bivariate binary case (see, for instance, Radice et al., 2016, and references therein) it is not necessary to penalize the correlation coefficient since the behavior of the respective log-likelihood function suggests that there is enough information that can be exploited in estimation. Moreover, while the analytical score vectors and Hessian matrices are readily available for bivariate binary models, they are not in the multivariate binary context.

This paper also illustrates the use of `SemiParTRIV()` in the package `SemiParBIVProbit` (Marra & Radice, 2017) for the `R` environment (Team, 2016), which implements the advances discussed in this paper. Current functions for fitting trivariate probit models are `triprobit()` (Terracol, 2002) or `mvprobit()` (Cappellari & Jenkins, 2003) in `STATA` (LP, 2015), and `mvProbit()` in the `R mvProbit` package (Henningsen, 2015). These implementations do not deal with the problems that this paper addresses. Moreover, `mvProbit()` may be unusably slow (as the author points out) and it requires all equations to have the same set of covariates. Note that we have focused on trivariate binary models, however the formulation in Section 2 can in principle be extended to the multivariate case as is the proposed estimation framework (see, for instance, the lemma and propositions in Sections 3 and 4).

The remainder of the paper is organised as follows. Section 2 introduces the trivariate probit model with additive predictors. Section 3 provides details on the model's likelihood whereas Section 4 discusses parameter estimation. Section 5 extends the estimation method by introducing a correlation-based penalty approach. Section 6 illustrates the proposed method through a case study whose aim is to estimate a model for three binary outcomes of newborn infants in North Carolina. The last section summarizes the paper and discusses some possibilities for future research. The supplementary on-line contains various details and proofs.

2. TRIVARIATE PROBIT MODEL WITH FLEXIBLE COVARIATE EFFECTS

The aim of the paper is to estimate and to make inference from a trivariate binary model in which the responses are determined by

$$y^*_{mi} = \mathbf{v}^\top_{mi}\boldsymbol{\gamma}_m + \sum_{\nu_m=1}^{\tilde{N}_m} s_{m\nu_m}(z_{m\nu_m i}) + \varepsilon_{mi}, \;\; i = 1, \ldots, n, \;\; \forall m = 1, 2, 3, \tag{2.1}$$

where n is the sample size, y^*_{mi} is a latent continuous variable, $\mathbf{v}_{mi}$ contains binary and/or categorical predictors, vector $\boldsymbol{\gamma}_m$ represents the effects of the variables in $\mathbf{v}_{mi}$, and $s_{m\nu_m}(z_{m\nu_m i})$ is a smooth function of continuous covariate $z_{m\nu_m i}, \forall \nu_m = 1, \ldots, \tilde{N}_m$ with $\tilde{N}_m$ being the number of smooth terms in the m^{th} equation. Latent variable y^*_{mi} determines the observed outcome as follows: if $y^*_{mi} > 0$ then $y_{mi} = 1$ and 0 otherwise. As for the error terms, we have that $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})^\top \overset{iid}{\sim} \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \vartheta_{12} & \vartheta_{13} \\ \vartheta_{21} & 1 & \vartheta_{23} \\ \vartheta_{31} & \vartheta_{32} & 1 \end{pmatrix}.$$

The error variances in $\boldsymbol{\Sigma}$ are normalized to unity (e.g., Greene, 2003, pp. 728), while the off-diagonal elements represent the correlations between the error terms and $\vartheta_{kz} = \vartheta_{zk}$ for $z \neq k$.

Smooth functions can be specified in several ways; see Ruppert et al. (2003) for details. We opted for the regression spline approach popularized by Eilers & Marx (1996) because of its computational efficiency, theoretical properties and flexibility in representing several types of covariate effects (e.g., Wood, 2006; Yoshida & Naito, 2014). Using this approach, $s_{m\nu_m}(z_{m\nu_m i})$ is approximated by a linear combination of known basis functions $b_{m\nu_m j}(z_{m\nu_m i})$ and regression parameters $\alpha_{m\nu_m j}$. That is,

$$s_{m\nu_m}(z_{m\nu_m i}) \approx \sum_{j=1}^{J_{m\nu_m}} \alpha_{m\nu_m, j} b_{m\nu_m, j}(z_{m\nu_m i}) = \mathbf{L}_{m\nu_m}(z_{m\nu_m i})\boldsymbol{\alpha}_{m\nu_m}, \tag{2.2}$$

where $\mathbf{L}_{m\nu_m}(z_{m\nu_m i})$ is a vector containing the $J_{m\nu_m}$ basis functions evaluated at $z_{m\nu_m i}$, that is $\mathbf{L}_{m\nu_m}(z_{m\nu_m i}) = \{b_{m\nu_m,1}(z_{m\nu_m i}), b_{m\nu_m,2}(z_{m\nu_m i}), \ldots, b_{m\nu_m, J_{m\nu_m}}(z_{m\nu_m i})\}$, and $\boldsymbol{\alpha}_{m\nu_m}$ is the corresponding parameter vector defined as $\boldsymbol{\alpha}_{m\nu_m} = \left(\alpha_{m\nu_m,1}, \alpha_{m\nu_m,2}, \ldots, \alpha_{m\nu_m, J_{m\nu_m}}\right)^\top, \forall m, \nu_m$. Moreover, each $\boldsymbol{\alpha}_{m\nu_m}$ has an associated quadratic penalty $\lambda_{m\nu_m}\boldsymbol{\alpha}^\top_{m\nu_m}\mathbf{S}_{m\nu_m}\boldsymbol{\alpha}_{m\nu_m}$ which enforces specific properties on the $m\nu^{th}_m$ function, such as smoothness. Smoothing parameter $\lambda_{m\nu_m} \in [0, \infty)$ controls the trade-off between fit

and smoothness. The overall penalty can be written as $\boldsymbol{\alpha}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3)^\top$, $\boldsymbol{\alpha}_m^\top = \left(\boldsymbol{\alpha}_{m1}^\top, \ldots, \boldsymbol{\alpha}_{m\tilde{N}_m}^\top\right)$ $\forall m$, $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{m=1}^{3} \sum_{\nu_m=1}^{\tilde{N}_m} \lambda_{m\nu_m} \mathbf{S}_{m\nu_m}$, $\boldsymbol{\lambda}$ is a vector containing all smoothing parameters and $\mathbf{S}_{m\nu_m}$ are positive definite or semi-definite symmetric known square matrices. Centering constraint $\sum_i s_{m\nu_m}(z_{m\nu_m i}) = 0$ is imposed on all smooth terms in the model for identification purposes. The above formulation allows us to represent many types of covariate effects depending on the nature of the covariate(s) considered; some common examples are described in Supplementary Material A.

Using regression spline representation (2.2), we can express (2.1) in a more compact way as

$$y_{mi}^* = \mathbf{v}_{mi}^\top \boldsymbol{\gamma}_m + \mathbf{L}_{mi}^\top \boldsymbol{\alpha}_m + \varepsilon_{mi} = \eta_{mi} + \varepsilon_{mi},$$

where $\eta_{mi} = \mathbf{v}_{mi}^\top \boldsymbol{\gamma}_m + \mathbf{L}_{mi}^\top \boldsymbol{\alpha}_m = \left(\mathbf{v}_{mi}^\top, \mathbf{L}_{mi}^\top\right)(\boldsymbol{\gamma}_m, \boldsymbol{\alpha}_m)^\top = \mathbf{x}_{mi}^\top \boldsymbol{\beta}_m$ and $\mathbf{L}_{mi}^\top = \{\mathbf{L}_{m1}(z_{m1i})^\top, \ldots, \mathbf{L}_{m\tilde{N}_m}(z_{m\tilde{N}_m i})^\top\}$, where $\mathbf{x}_{mi}$ and $\boldsymbol{\beta}_m$ are vectors of length P_m.

3. MODEL'S LIKELIHOOD

Because of the presence of additive predictors in the model, classical maximum likelihood estimation (MLE) is not appropriate for parameter estimation as over-fitting is likely to occur in practical situations. This issue is overcome by adopting a penalized approach where a penalty term, controlling for the model's smoothness, is added to the original objective function. Simultaneous estimation of all parameters of the trivariate additive probit model is therefore achieved by penalized MLE (PMLE) through problem

$$\hat{\boldsymbol{\delta}} := \arg\min_{\boldsymbol{\delta}} -\ell_p(\boldsymbol{\delta}) = \arg\min_{\boldsymbol{\delta}} -\{\log \mathcal{L}(\mathbf{Y}; \boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\alpha}\}, \tag{3.3}$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}^\top, \boldsymbol{\vartheta}^\top)^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top)^\top$, $\boldsymbol{\vartheta} = (\vartheta_{12}, \vartheta_{13}, \vartheta_{23})^\top$, $\boldsymbol{\alpha}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\alpha} = \boldsymbol{\delta}^\top \tilde{\mathbf{S}}_{\boldsymbol{\lambda}} \boldsymbol{\delta}$ with

$\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$=diag $\left(\mathbf{0}_{\tilde{P}_1}^\top, \lambda_{1\nu_1}\mathbf{S}_{1\nu_1}, \ldots, \lambda_{1\tilde{N}_1}\mathbf{S}_{1\tilde{N}_1}, \mathbf{0}_{\tilde{P}_2}^\top, \lambda_{2\nu_2}\mathbf{S}_{2\nu_2}, \ldots, \lambda_{2\tilde{N}_2}\mathbf{S}_{2\tilde{N}_2}, \mathbf{0}_{\tilde{P}_3}^\top, \lambda_{3\nu_3}\mathbf{S}_{3\nu_3}, \ldots, \lambda_{3\tilde{N}_3}\mathbf{S}_{3\tilde{N}_3}, 0, 0, 0\right)$,

$\mathbf{0}_{\tilde{P}_m}^\top = \left(0_{m1}, \ldots, 0_{m\tilde{P}_m}\right)$ and $\tilde{P}_m$ denotes the number of variables in $\mathbf{v}_m$. For a 3-dimensional binary response vector we have 2^3 trivariate probabilities expressed via the cumulative distribution function (cdf)

of the trivariate normal distribution. The likelihood is given by the joint density of observed outcomes

$$\mathcal{L}(\mathbf{Y};\boldsymbol{\delta}) = \prod_{i=1}^{n}\prod_{\tilde{k}=1}^{2^3}\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \prod_{i=1}^{n}\prod_{\tilde{k}=1}^{2^3}\boldsymbol{\Psi}_{i\tilde{k}}^{\mathcal{Y}_{i\tilde{k}}},$$

where $\mathcal{L}_{i\tilde{k}}$ is derived from Lemma 3.1 for $M = 3$. Term $\mathcal{Y}_{i\tilde{k}}$ denotes an indicator variable for the $\tilde{k}^{th}$ combination of the three possible events $y_{1i} = \bar{e}_1, y_{2i} = \bar{e}_2, y_{3i} = \bar{e}_3$ with $\bar{e}_m \in \{0, 1\}\ \forall m$ and $\boldsymbol{\Psi}_{i\tilde{k}}$ is the corresponding trivariate normal cdf. For instance, if $\tilde{k} = 3$ corresponds to events $y_{1i} = y_{3i} = 1$ and $y_{2i} = 0$ then $\mathcal{Y}_{i3} = y_{1i}(1 - y_{2i})y_{3i}$ and $\boldsymbol{\Psi}_{i3} = \mathbb{P}(y_{1i} = 1, y_{2i} = 0, y_{3i} = 1)$.

LEMMA 3.1 Quantity $\mathcal{L}_{i\tilde{k}}$ evaluated at the vector $(\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i)_{\tilde{k}}$ is equal to the cdf of a multivariate standardized normal vector with correlation matrix $(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)_{\tilde{k}}$, that is

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \boldsymbol{\Psi}_{i\tilde{k}}^{\mathcal{Y}_{i\tilde{k}}} = \left\{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)_{\tilde{k}})\right\}^{\mathcal{Y}_{i\tilde{k}}} = \left\{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{w}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i)_{\tilde{k}})\right\}^{\mathcal{Y}_{i\tilde{k}}},$$

where $\boldsymbol{w}_i = \boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i = (w_{1,i}, w_{2,i}, \ldots, w_{M,i})^\top$, $\boldsymbol{\Upsilon}_i = \boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i$, $w_{m,i} = \tilde{y}_{mi}\eta_{mi}$, for $\tilde{y}_{mi} = (2y_{mi} - 1)$, $\eta_{mi} = \mathbf{x}_{mi}^\top\boldsymbol{\beta}_m$, $\boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i}, \ldots, \eta_{Mi})^\top$, $\boldsymbol{\mathcal{B}}_i$ denotes a diagonal $M \times M$ matrix with main diagonal elements $\tilde{y}_{mi} = (2y_{mi} - 1)$ that depend on y^*_{mi}, that is $\boldsymbol{\mathcal{B}}_i = \text{diag}(2y_{1i} - 1, 2y_{2i} - 1, \ldots, 2y_{Mi} - 1)$.

Proof. See Supplementary Material B. □

We can therefore express the log-likelihood function as

$$\log\mathcal{L}(\mathbf{Y};\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) = \sum_{i=1}^{n}\sum_{\tilde{k}=1}^{2^3}\ell_{i\tilde{k}}(\boldsymbol{\delta}) = \sum_{i=1}^{n}\sum_{\tilde{k}=1}^{4}\left\{\mathcal{Y}_{i\tilde{k}}\log\boldsymbol{\Psi}_{i\tilde{k}} + \mathcal{Y}_{i(4+\tilde{k})}\log\boldsymbol{\Psi}_{i(4+\tilde{k})}\right\},$$

where $\boldsymbol{\Psi}_{i\tilde{k}} = \Phi_{3,\boldsymbol{\varepsilon}_i}((\boldsymbol{w}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i)_{\tilde{k}})$, $\boldsymbol{\Psi}_{i(4+\tilde{k})} = \Phi_{3,\boldsymbol{\varepsilon}_i}(-(\boldsymbol{w}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i)_{\tilde{k}})$, $\Phi_{3,\boldsymbol{\varepsilon}_i}$ corresponds to trivariate normal integrals, and $\boldsymbol{w}_i$ and $\boldsymbol{\Upsilon}_i$ are defined in Lemma 1. Note that for each $\tilde{k}$ the form of $\boldsymbol{w}_i$ and $\boldsymbol{\Upsilon}_i$ is different as their structure depends on the $\tilde{k}^{th}$ combination of the three possible events. In general there are no exact methods for calculating the multivariate normal (MVN) probabilities $\Phi_{M,\boldsymbol{\varepsilon}_i}$, for $M \geqslant 2$. Accurate approximations, however, can be obtained via the `R` function `pmnorm()` in package `mnormt` (Azzalini, 2014). The approximation method by Trinh & Genz (2015) is another possibility for computing the MVN probabilities. As compared to `pmnorm()`, this approach gains computational

speed but becomes less accurate for highly correlated responses. Both methods have been implemented in `SemiParTRIV()`; their full description can be found in Supplementary Materials C.1 and C.2. Once $\mathbb{P}(y_{1i} = 1, y_{2i} = 1, y_{3i} = 1)\ \forall i$ has been obtained, the remaining probabilities can be efficiently calculated using relationship $\sum_{i=1}^{n} \{p_{111i} + p_{110i} + p_{101i} + p_{011i} + p_{000i} + p_{001i} + p_{010i} + p_{100i}\} = \sum_{i=1}^{n} \{p_{11i} + p_{10i} + p_{01i} + p_{00i}\} = \sum_{i=1}^{n} \{p_{1i} + p_{0i}\} = 1$, where $p_{\bar{e}_1\bar{e}_2\bar{e}_3 i} = \mathbb{P}(y_{1i} = \bar{e}_1, y_{2i} = \bar{e}_2, y_{3i} = \bar{e}_3)$, $p_{\bar{e}_1\bar{e}_2 i} = \mathbb{P}(y_{1i} = \bar{e}_1, y_{2i} = \bar{e}_2)$ and $p_{\bar{e}_1 i} = \mathbb{P}(y_{1i} = \bar{e}_1)$.

Since the correlation parameters can only take values in $[-1, 1]$, we use Fisher transformation $\vartheta^*_{zk} = \tanh^{-1}(\vartheta_{zk})$ and redefine $\boldsymbol{\delta}$ as $(\boldsymbol{\beta}^\top, \boldsymbol{\vartheta}^{*\top})^\top$ to ensure that in optimization $\boldsymbol{\delta} \in \mathbb{R}^Q$, where $\boldsymbol{\vartheta}^* = (\vartheta^*_{12}, \vartheta^*_{13}, \vartheta^*_{23})^\top$ and Q is the total number of parameters in $\boldsymbol{\delta}$. To ensure positive-definiteness of $\boldsymbol{\Sigma}$, we need to include range restrictions on the correlations; in this case, if we fix ϑ_{13} and ϑ_{23} then ϑ_{12} is restricted to take values in $\left(\vartheta_{13}\vartheta_{23} - \sqrt{(1-\vartheta^2_{13})(1-\vartheta^2_{23})}, \vartheta_{13}\vartheta_{23} + \sqrt{(1-\vartheta^2_{13})(1-\vartheta^2_{23})}\right)$. In practice, such a restriction is imposed using the eigenvalue method (Rousseeuw & Molenberghs, 1993). A detailed description of the approach and the relevant geometric proof can be found in Supplementary Materials D.1 and D.2 for reader's convenience.

4. Parameter estimation

Joint estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ via (3.3) would clearly lead to severe over-fitting as the optimal value of $\ell_p(\boldsymbol{\delta})$ would be reached when $\hat{\boldsymbol{\lambda}} = \mathbf{0}$ (e.g., Ruppert et al., 2003). Following Gu (2002), Marra et al. (2016) and Wood (2004), we estimate the model and smoothing parameters using a two stage approach; one step concerns estimation of $\boldsymbol{\delta}$ conditional on $\boldsymbol{\lambda}$ and the other estimation of $\boldsymbol{\lambda}$ conditional on $\boldsymbol{\delta}$. Note that such an approach is philosophically very similar to the Bayesian estimation method discussed, for instance, by Klein & Kneib (2016) where Bayesian sampling is used to estimate $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ conditional on each other.

Holding $\boldsymbol{\lambda}$ fixed at a vector of values, we seek to minimize $-\ell_p(\boldsymbol{\delta})$. This is achieved via a trust-region algorithm which has generally proved to be more stable and faster than standard numerical optimization procedures for simultaneous models (e.g., Donat & Marra, 2016; Radice et al., 2016). Each iteration $\varkappa$ of

the trust-region algorithm solves the sub-problem

$$\min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]}) := -\left\{ \ell_p\left(\boldsymbol{\delta}^{[\varkappa]}\right) + \boldsymbol{s}^\top \boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa]}) + \frac{1}{2}\boldsymbol{s}^\top \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\varkappa]})\boldsymbol{s} \right\}$$
$$\text{subject to } \|\boldsymbol{s}\| \leqslant \boldsymbol{\Delta}^{[\varkappa]},$$
$$\boldsymbol{\delta}^{[\varkappa+1]} = \arg\min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]}) + \boldsymbol{\delta}^{[\varkappa]},$$

where $\mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]})$ is a quadratic approximation of ℓ_p at $\boldsymbol{\delta}^{[\varkappa]}$, $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa]})$ denotes the penalized score function defined as $\boldsymbol{g}(\boldsymbol{\delta}^{[\varkappa]}) - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[\varkappa]}$, $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\varkappa]})$, the penalized Hessian matrix, is given by $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}^{[\varkappa]}) - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}$, $\|\cdot\|$ denotes the Euclidean norm and $\boldsymbol{\Delta}^{[\varkappa]}$ is the radius of the trust region. A more detailed description of the trust region approach is given in Supplementary Material E. The analytical score function, $\boldsymbol{g}_i(\boldsymbol{\delta}^{[\varkappa]}) = \nabla_{\boldsymbol{\delta}}\ell_i\left(\boldsymbol{\delta}^{[\varkappa]}\right)$, and Hessian matrix, $\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}^{[\varkappa]}) = \nabla_{\boldsymbol{\delta}}\nabla_{\boldsymbol{\delta}}^\top\ell_i\left(\boldsymbol{\delta}^{[\varkappa]}\right)$, required to implement the trust-region approach are computed using

$$\nabla_{\boldsymbol{\delta}}\ell_i(\boldsymbol{\delta}) = \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i} = \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\right\}, \tag{4.4}$$

$$\nabla_{\boldsymbol{\delta}}\nabla_{\boldsymbol{\delta}}^\top\ell_i(\boldsymbol{\delta}) = \frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i}\frac{\partial^2\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} + \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \frac{\partial^2\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i\partial\bar{\boldsymbol{\eta}}_i^\top}\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}} = \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\right\}\frac{\partial^2\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} + \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{-\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}\boldsymbol{\Psi}_{i\tilde{k}}^\top}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\left(\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i^\top}\right)^\top + \frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial^2\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i\partial\bar{\boldsymbol{\eta}}_i}^\top\right\}\left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right), \tag{4.5}$$

where, for each i, $\bar{\boldsymbol{\eta}}_i = (\eta_{1i}, \eta_{2i}, \eta_{3i}, \eta_{4i}, \eta_{5i}, \eta_{6i})^\top$ with $(\eta_{4i}, \eta_{5i}, \eta_{6i}) = (\vartheta_{12}^*, \vartheta_{13}^*, \vartheta_{23}^*)$, $\partial\bar{\boldsymbol{\eta}}_i/\partial\boldsymbol{\delta} = \text{diag}\left(\partial\eta_{1i}/\partial\boldsymbol{\beta}_1, \partial\eta_{2i}/\partial\boldsymbol{\beta}_2, \partial\eta_{3i}/\partial\boldsymbol{\beta}_3, \partial\vartheta_{12}^*/\partial\vartheta_{12}^*, \partial\vartheta_{13}^*/\partial\vartheta_{13}^*, \partial\vartheta_{23}^*/\partial\vartheta_{23}^*\right) = \text{diag}\left(\partial\eta_{1i}/\partial\boldsymbol{\beta}_1, \partial\eta_{2i}/\partial\boldsymbol{\beta}_2, \partial\eta_{3i}/\partial\boldsymbol{\beta}_3, 1, 1, 1\right)$ and $\partial\ell(\boldsymbol{\delta})/\partial\bar{\boldsymbol{\eta}}_i$=$(\partial\ell(\boldsymbol{\delta})/\partial\eta_{1i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{2i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{3i}, \partial\ell(\boldsymbol{\delta})/\partial\vartheta_{12}^*, \partial\ell(\boldsymbol{\delta})/\partial\vartheta_{13}^*, \partial\ell(\boldsymbol{\delta})/\partial\vartheta_{23}^*)^\top$. Predictor $\bar{\boldsymbol{\eta}}_i$ is functionally dependent on the Q-vector $\boldsymbol{\delta}$, that is $\bar{\boldsymbol{\eta}}_i = \bar{\boldsymbol{\eta}}_i(\boldsymbol{\delta})$. The difficulty with deriving analytical expressions for the derivative components in (4.4) and (4.5) is that they require working with trivariate integrals, which is not straightforward. Propositions 1 and 2 provide the key derivatives for the log-likelihood function of a generic multivariate probit model with correlation matrix structured as

$$\boldsymbol{\Upsilon}_i^* = \begin{pmatrix} 1 & r_{12,i}^* & r_{13,i}^* & \cdots & r_{1M,i}^* \\ r_{12,i}^* & 1 & r_{23,i}^* & \cdots & r_{2M,i}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1M,i}^* & r_{2M,i}^* & r_{3M,i}^* & \cdots & 1 \end{pmatrix},$$

where $r^*_{zk,i} = \tanh(\vartheta^*_{zk})(2y_{zi}-1)(2y_{ki}-1), \forall z,k,i$. The propositions below have been used to implement expressions (4.4) and (4.5) after setting $M=3$.

PROPOSITION 1 Assume that $\boldsymbol{w}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}^*_i$. Then the first-order derivative of the M-variate normal cdf $\Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)$ with respect to $\boldsymbol{\beta}_m, \forall m = 1,\ldots,M$, can be expressed as

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)}{\partial \boldsymbol{\beta}_m} = \phi(w_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{w}_{-m,i}|w_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i)(2y_{mi}-1)\mathbf{x}^\top_{mi},$$

where M denotes the total number of equations under a multivariate probit framework, $w_{m,i}$ denotes the linear predictor of the m^{th} equation and is equal to $(2y_{mi}-1)\mathbf{x}^\top_{mi}\boldsymbol{\beta}_m$, $\boldsymbol{\beta}_m$ denotes the parameter vector of covariate vector $\mathbf{x}_{mi}$ and the vector of linear predictors $\boldsymbol{w}_{-m,i}$ is defined as $(w_{1,i}, w_{2,i}, \ldots, w_{m-1,i}, w_{m+1,i}, \ldots, w_{M,i})^\top$. The mean $\boldsymbol{M}^{*m}_i$ and variance-covariance matrix $\boldsymbol{\Theta}^{*m}_i$ is equal to $\boldsymbol{\Theta}^{*m}_{21,i} w_{m,i}$ and $\boldsymbol{\Theta}^{*m}_{22,i} - \boldsymbol{\Theta}^{*m}_{21,i}\boldsymbol{\Theta}^{*m}_{12,i}$, respectively, with $\boldsymbol{\Theta}^{*m}_{12,i}$, $\boldsymbol{\Theta}^{*m}_{21,i}$ and $\boldsymbol{\Theta}^{*m}_{22,i}$ defined by re-ordering $\boldsymbol{\Upsilon}^*_i$ as follows

$$\boldsymbol{\Upsilon}^{*m}_i = \left(\begin{array}{c|c} \overbrace{\Theta^{*m}_{11,i}}^{1\times 1} & \overbrace{\boldsymbol{\Theta}^{*m}_{12,i}}^{1\times(M-1)} \\ \hline \underbrace{\boldsymbol{\Theta}^{*m}_{21,i}}_{(M-1)\times 1} & \underbrace{\boldsymbol{\Theta}^{*m}_{22,i}}_{(M-1)\times(M-1)} \end{array} \right).$$

The element $\Theta^{*m}_{11,i}$ is equal to 1, the off-diagonal blocks $\boldsymbol{\Theta}^{*m}_{12,i}$ and $\boldsymbol{\Theta}^{*m}_{21,i}$ consist of the correlations $r^*_{m\varpi,i}$, $\forall\, \varpi \in \{1:M\} \setminus m$, for $m \neq \varpi$ and the symmetric sub-matrix $\boldsymbol{\Theta}^{*m}_{22,i}$ has main diagonal elements equal to 1 and off-diagonals equal to $r^*_{\bar{\varphi}\varpi,i}, \forall \bar{\varphi}, \varpi \in \{1:M\} \setminus m$, for $\bar{\varphi} \neq \varpi$.

Proof. See Supplementary Material F.1. □

PROPOSITION 2 Assume that $\boldsymbol{w}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}^*_i$. Then the first-order derivative of the M-variate normal cdf $\Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)$ with respect to $\vartheta^*_{zk}, \forall z = 1,\ldots,M-1, k = z+1,\ldots M$, can be expressed as

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)}{\partial \vartheta^*_{zk}} = \phi_2(\boldsymbol{w}_{zk,i}; \mathbf{0}, \boldsymbol{\Theta}^{*zk}_i)\Phi_{M-2}(\boldsymbol{w}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}^{*-zk}_i, \boldsymbol{\Theta}^{*-zk}_i) \times$$
$$(2y_{zi}-1)(2y_{ki}-1)\frac{4e^{2\vartheta^*_{zk}}}{(e^{2\vartheta^*_{zk}}+1)^2},$$

where M denotes the total number of equations under a multivariate probit framework, $\boldsymbol{w}_{zk,i} = (w_{z,i}, w_{k,i})^\top$, $w_{z,i}$ and $w_{k,i}$ refer to the linear predictors of the z^{th} and k^{th} equations respectively and are equal to $(2y_{mi} - 1)\mathbf{x}_{mi}^\top \boldsymbol{\beta}_m$, $\forall m = z, k$, and $\boldsymbol{\beta}_m$ denotes the parameter vector of covariate vector $\mathbf{x}_{mi}$. The vector of linear predictors $\boldsymbol{w}_{-zk,i}$ is defined as $(w_{1,i}, w_{2,i}, \ldots, w_{z-1,i}, w_{z+1,i}, \ldots, w_{k-1,i}, w_{k+1,i}, \ldots w_{M,i})^\top$, while parameter $\vartheta^*_{zk} = \tanh^{-1}(\vartheta_{zk})$ where ϑ_{zk} denotes the correlation coefficient between the z^{th} and k^{th} responses. The variance-covariance matrix $\boldsymbol{\Theta}_i^{*zk}$ is equal to $\boldsymbol{\Theta}_{11,i}^{*zk}$, while the mean $\boldsymbol{M}_i^{*-zk}$ and variance-covariance matrix $\boldsymbol{\Theta}_i^{*-zk}$ is equal to $\boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{w}_{zk}$ and $\boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}$, respectively. The sub-matrices $\boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{\Theta}_{12,i}^{*zk}$, $\boldsymbol{\Theta}_{21,i}^{*zk}$ and $\boldsymbol{\Theta}_{22,i}^{*zk}$ are defined by re-ordering $\boldsymbol{\Upsilon}_i^*$ as follows

$$\boldsymbol{\Upsilon}_i^{*zk} = \left(\cdots \overbrace{\underbrace{\boldsymbol{\Theta}_{11,i}^{*zk}}_{(M-2)\times 2}}^{2\times 2} \vdots \cdots \overbrace{\underbrace{\boldsymbol{\Theta}_{12,i}^{*zk}}_{(M-2)\times(M-2)}}^{2\times(M-2)} \cdots \right).$$

The sub-matrix $\boldsymbol{\Theta}_{11,i}^{*zk}$ has unit diagonals and off-diagonals equal to $r^*_{zk,i}$. The first row (column) of $\Theta_{12,i}^{*zk}$ ($\Theta_{21,i}^{*zk}$) contains the correlations $r^*_{z\bar{\varrho},i}$, for $\bar{\varrho} \in \{1:M\} \backslash z$, while the second row (column) of $\Theta_{12,i}^{*zk}$ ($\Theta_{21,i}^{*zk}$) contains the correlations $r^*_{\bar{v}k,i}$, for $\bar{v} \in \{1:M\} \setminus k$. The diagonal block $\boldsymbol{\Theta}_{22,i}^{*zk}$ is a symmetric matrix with unit diagonals and off-diagonal elements equal to $r^*_{\chi\psi,i}$, $\forall\, \chi, \psi \in \{1:M\} \setminus \{z,k\}$ for $\chi \neq \psi$.

Proof. See Supplementary Material F.2. □

The analytical derivatives have been verified via numerical differentiation using the `R` package `numDeriv` (Gilbert & Varadhan, 2015). Full matrices $\boldsymbol{\Upsilon}_i^{*m}$ and $\boldsymbol{\Upsilon}_i^{*zk}$ can be found in Supplementary Material G.

Estimation of $\boldsymbol{\lambda}$ conditional on an updated estimate for $\boldsymbol{\delta}$ is obtained using the method detailed in Supplementary Material H for the sake of space. The two steps are iterated until the algorithm satisfies criterion $\frac{\left|\ell(\boldsymbol{\delta}^{[\varkappa+1]}) - \ell(\boldsymbol{\delta}^{[\varkappa]})\right|}{0.1 + \left|\ell(\boldsymbol{\delta}^{[\varkappa+1]})\right|} < 1e-07$. At convergence, well founded point-wise confidence intervals for linear and non-linear functions of the model coefficients can be obtained using result $\boldsymbol{\delta} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1}\right)$; see Supplementary Material H.1 for further details.

4.1 *Simulation Study I*

A simulation study was conducted to investigate the practical performance of the proposed approach as compared to the available alternative routine `mvprobit()` available in `STATA`.

4.1.1 *DGP1* In order to compare the results obtained from `SemiParTRIV()` and `mvprobit()`, we employed a Data Generating Process (DGP) based on the fully parametric model. Exact simulation settings and the code used to generate the data can be found in Supplementary Material I.1. The syntax used to fit trivariate probit models is

```
out <- SemiParTRIV(formula = f.l, data = dat)
```

where `f.l` consists of a list of three equations

```
eqn1 <- y1 ~ v1 + z1; eqn2 <- y2 ~ v1 + z1; eqn3 <- y3 ~ v1 + z1
f.l <- list(eqn1, eqn2, eqn3)
```

and `v1` and `z1` denote the binary and continuous covariates, respectively. Argument `data` refers to the data frame containing the variables in the model.

The results for $n = 1000$ are summarized in Figure 1, whereas those for $n = 10000$ can be found in Supplementary Material I.1. The regression coefficient estimates of both methods are satisfactory and converge to their true values as n increases. As expected, the variability of the estimates decreases as the sample size grows large. As for the correlation parameters, `SemiParTRIV()` considerably outperforms `mvprobit()` whose estimates do not improve as n increases. This may have important inferential implications; for instance, obtaining unbiased joint outcome probabilities requires accurate estimation of the correlation coefficients (e.g., Neelon et al., 2014). For the sake of space, a brief discussion regarding the unsatisfactory performance of `mvprobit()` is reported in Supplementary Material I.1, while the `STATA` and `R` codes used to run the models for the above study are given in Supplementary Material I.1.1.

4.1.2 *DGP2* The proposed approach does have some limitations, however. On occasion, the algorithm does not satisfy the first and second order necessary conditions for convergence (that is, zero gradient and positive definite Hessian matrix). When this occurs, we observed that the non-zero gradient components and/or negative eigenvalues of the Hessian matrix are typically associated with the correlation parameters. To shed light on this issue, we conducted more simulation studies based on different configurations of the correlation matrix. We refer to the simulation settings of one such study as DGP2 whose description is given in Supplementary Material I.1. Table 1 displays the percentage biases and root mean squared errors (RMSEs) for the correlation estimates. The results show that the estimation performance improves as n grows large, however at $n = 1000$ the method is not deemed to perform satisfactorily. The estimated regression coefficients (not shown here) were similar to those of the previous study at both sample sizes. The R code used for this study is given in Supplementary Material I.1.2.

To gain more insights into the above mentioned issue, we looked at the log-likelihood behavior over the correlation parameters. For instance, we produced univariate transects through ℓ by evaluating $\ell(\boldsymbol{\delta})$ at the optimal MLE values for $\boldsymbol{\beta}$, ϑ^*_{12} and ϑ^*_{13}, for a grid of ϑ^*_{23} values. Figure 2 shows the corresponding $\ell(\boldsymbol{\delta})$ versus ϑ^*_{23}, based on 10 replicates, from which we observe a minimum that tends to be very shallow. This suggests that at small sample sizes the log-likelihood (and thus the model) may provide little information with which one can make inferences. Greater uncertainty is also expected. When this happens the parameter is weakly or not identified. The methodology described in the next section addresses this issue.

5. CORRELATION-BASED PENALTY

The aim of this section is to further augment the penalized log-likelihood function by introducing a penalty which addresses the difficulty in estimating the correlation parameters. The PMLE problem (3.3) then becomes

$$\hat{\boldsymbol{\delta}} := \arg\min_{\boldsymbol{\delta}} -\{\ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\top \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta} - \mathcal{P}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})\}, \tag{5.6}$$

where $\mathcal{P}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ is a penalty acting on the correlations that depends on λ_{ϑ^*} which determines the amount of shrinkage required for ϑ^*_{zk}, $\forall z, k$. In this work, we employ the Ridge, Lasso and Adaptive Lasso approaches.

Suppose that $\mathbf{R}_q = \text{diag}\,(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ where the value of 1 on the $(q,q)^{th}$ entry of the matrix corresponds to the q^{th} parameter in $\boldsymbol{\delta}$, $\forall q = 1, \ldots, Q$. Then, the penalties can be expressed as

$$\textbf{Lasso: } \mathcal{P}^{\text{L}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathcal{P}^{\text{L}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) = \lambda_{\vartheta^*}\left(|\vartheta^*_{12}| + |\vartheta^*_{13}| + |\vartheta^*_{23}|\right), \tag{5.7}$$

$$\textbf{Ridge: } \mathcal{P}^{\text{R}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathcal{P}^{\text{R}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_2^2) = \frac{1}{2}\lambda_{\vartheta^*}\left(\vartheta^{*2}_{12} + \vartheta^{*2}_{13} + \vartheta^{*2}_{23}\right), \tag{5.8}$$

$$\textbf{Ad. Lasso: } \mathcal{P}^{\text{AL}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathcal{P}^{\text{AL}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) = \lambda_{\vartheta^*}\left(\frac{|\vartheta^*_{12}|}{|\hat{\vartheta}^{*\text{MLE}}_{12}|^{\bar{\gamma}}} + \frac{|\vartheta^*_{13}|}{|\hat{\vartheta}^{*\text{MLE}}_{13}|^{\bar{\gamma}}} + \frac{|\vartheta^*_{23}|}{|\hat{\vartheta}^{*\text{MLE}}_{23}|^{\bar{\gamma}}}\right), \tag{5.9}$$

$\forall q = Q-2, Q-1, Q$, where superscripts L, R, and AL refer to the Lasso, Ridge and Adaptive Lasso penalties, respectively. The expression for the Adaptive Lasso is obtained as follows. Suppose that $\hat{\boldsymbol{\delta}}$ is a root-n-consistent estimator for $\boldsymbol{\delta}$, in which case we can use $\hat{\boldsymbol{\delta}}^{\text{MLE}}$. Then by picking a $\bar{\gamma} > 0$ it is possible to define adaptive weights as $\text{w}_q = 1/|\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\text{MLE}}|^{\bar{\gamma}}$ (Zou, 2006). Thus, we have that $\text{w}_{Q-2} = 1/|\hat{\vartheta}^{*\text{MLE}}_{12}|^{\bar{\gamma}}$, $\text{w}_{Q-1} = 1/|\hat{\vartheta}^{*\text{MLE}}_{13}|^{\bar{\gamma}}$ and $\text{w}_Q = 1/|\hat{\vartheta}^{*\text{MLE}}_{23}|^{\bar{\gamma}}$. Based on simulation studies, we found that $\bar{\gamma} = 1$ works well in most situations, however a sensitivity analysis trying different values for this parameter could be carried out. Note that when using Adaptive Lasso different amounts of shrinkage for each correlation are used and thus each coefficient is weighted differently. The derivation of expressions (5.7)-(5.9) can be found in Supplementary Material J.1.

5.1 *Computational aspects*

As pointed out by Ulbricht (2010), a penalty function should satisfy the following properties: (P.1) $\mathcal{P}_{\lambda_{\vartheta^*}}: \mathbb{R}^+ \to \mathbb{R}^+$, $\mathcal{P}_{\lambda_{\vartheta^*}}(\mathbf{0}) = \mathbf{0}$; (P.2) $\mathcal{P}_{\lambda_{\vartheta^*}}$ is continuous and strictly monotone in $\mathbf{R}_q^\top\boldsymbol{\delta}$; and (P.3) $\mathcal{P}_{\lambda_{\vartheta^*}}$ is continuously differentiable, $\forall \mathbf{R}_q\boldsymbol{\delta} \neq \mathbf{0}$, such that $\partial\mathcal{P}_{\lambda_{\vartheta^*}}/\partial\mathbf{R}_q\boldsymbol{\delta} > \mathbf{0}$. The Ridge penalty is a quadratic function and satisfies (P.1)-(P.3). By contrast, the Lasso and Adaptive Lasso penalties are singular at $\boldsymbol{\delta} = \mathbf{0}$ (and thus not differentiable at this point) and non-concave with respect to $\boldsymbol{\delta}$. In these cases, it would

be unfeasible to maximize the penalized likelihood function using the approach described in Section 3. We therefore elect to approximate these two non-differentiable penalties by differentiable ones. Such approximations are available in the literature. For instance, Fan & Li (2001) approximated quadratically the non-convex SCAD penalty, while Ulbricht (2010) applied this idea to Lasso penalties. Rippe et al. (2012) approximated quadratically the L_0-type penalty by employing a weighted Ridge penalty. In this work, we employ the local quadratic approximation approach.

5.1.1 *Approximations of non-differentiable norms* The non-differentiability of L_1-type penalties such as Lasso and Adaptive Lasso can be avoided by approximating a norm at the critical point $\|\mathbf{R}_q\boldsymbol{\delta}\|_1 = \mathbf{0}$. Let $\|\mathbf{R}_q\boldsymbol{\delta}\|_1 = \|\boldsymbol{\xi}_q\|_1$. As in Koch (1996), norm $\|\boldsymbol{\xi}_q\|_1$ in a penalty function can be approximated by $\left(\boldsymbol{\xi}_q^\top\boldsymbol{\xi}_q + \bar{c}\right)^{1/2}$, where $\bar{c}$ is a small positive real number which controls how close the approximation and the exact function are; Oelker & Tutz (2013) argue that $\bar{c} \approx 10^{-8}$ works well in most cases. Similarly as in Oelker & Tutz (2013), we combine this approximation with a trick by Fan & Li (2001) as well as an idea introduced by Ulbricht (2010). We assume that an approximation to each norm $\|\boldsymbol{\xi}_q\|_l$ exists such that $\|\boldsymbol{\xi}_q\|_l = \mathcal{K}_l(\boldsymbol{\xi}_q, \mathcal{C}) = \lim_{\bar{\mathcal{C}}\to\mathcal{C}} \mathcal{K}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$, where $\bar{\mathcal{C}}$ represents a set of possible tuning parameters, $\mathcal{C}$ is the set of boundary values for $\|\boldsymbol{\xi}_q\|_l$ and $\mathcal{K}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$ should be at least twice differentiable $\forall l \geqslant 1$. Additionally, for all $\boldsymbol{\xi}_q$, for which derivative $\partial \|\boldsymbol{\xi}_q\|_l / \partial \boldsymbol{\xi}_q$ is defined, we assume that $\partial \|\boldsymbol{\xi}_q\|_l / \partial \boldsymbol{\xi}_q = \lim_{\bar{\mathcal{C}}\to\mathcal{C}} \mathcal{D}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$, where $\mathcal{D}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}}) = \partial\mathcal{K}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}})/\partial\boldsymbol{\xi}_q \; \forall l$. We further assume that $\mathcal{D}_l(\mathbf{0}, \bar{\mathcal{C}}) = \mathbf{0}$. As mentioned above, the L_1 norm is approximated by $\mathcal{K}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}}) = (\boldsymbol{\xi}_q^\top\boldsymbol{\xi}_q + \bar{c})^{1/2}$. The first derivative $\mathcal{D}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}}) = \left(\boldsymbol{\xi}_q^\top\boldsymbol{\xi}_q + \bar{c}\right)^{-1/2}\boldsymbol{\xi}_q$ is a continuous approximation for the first-order derivative of the L_1 norm. In general, $\mathcal{K}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$ deviates only slightly from $\mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{C})$. That is, for $\boldsymbol{\xi}_q = \mathbf{0}$ the deviation is $\sqrt{\bar{c}}$, while for any other value of $\boldsymbol{\xi}_q$ the deviation is $< \sqrt{\bar{c}}$.

Penalty $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$, for $\mathcal{G} = \{\mathrm{L}, \mathrm{AL}\}$, can be locally approximated by a quadratic function as follows. Suppose that $\tilde{\boldsymbol{\delta}}$ is an initial value close to $\hat{\boldsymbol{\delta}}$. Then we approximate $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ by a Taylor expansion of order 1 at $\tilde{\boldsymbol{\delta}}$, that is, $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) \approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\tilde{\boldsymbol{\delta}}}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})^\top(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}})$. As proved in Supplementary Material J.2,

$\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ can be approximated as

$$\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) \approx \frac{1}{2}\boldsymbol{\delta}^\top \left\{ \nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top} \mathbf{R}_q\mathbf{R}_q^\top \right\} \boldsymbol{\delta} \approx \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}} \boldsymbol{\delta},$$

where $\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) = \partial \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) / \partial \|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1$, $\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}}) = \partial \|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1 / \partial \mathbf{R}_q\tilde{\boldsymbol{\delta}}$, $\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ has the following form

$$\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}} = \begin{pmatrix} \mathbf{0}_{Q\times Q} & \mathbf{0}_{Q\times 3} \\ \mathbf{0}_{3\times Q} & \mathbf{A}^{\mathcal{G}}_{\lambda_{\vartheta^*}} \end{pmatrix},$$

and $\mathbf{A}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ is a 3×3 diagonal matrix that corresponds to the correlation parameters that have to be penalized. The expressions for the penalty matrices of Lasso and Adaptive Lasso are

$$\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}} = \lambda_{\vartheta^*} \operatorname{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \right), \tag{5.10}$$

$$\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}} = \lambda_{\vartheta^*} \operatorname{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{12}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{13}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{23}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \right). \tag{5.11}$$

Note that $\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ needs to be updated at each iteration of the algorithm as it depends on the estimated correlations. In the Ridge penalty case we simply have $\boldsymbol{\Lambda}^{\mathrm{R}}_{\lambda_{\vartheta^*}} = \lambda_{\vartheta^*} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 1, 1, 1)$. The derivations of (5.10) and (5.11) are given in Supplementary Material J.3.

It follows that the penalized log-likelihood, score and Hessian matrix can be expressed as $\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}$, $\boldsymbol{g}_p(\boldsymbol{\delta}) = \boldsymbol{g}(\boldsymbol{\delta}) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}$ and $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\delta}) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}$, where $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}} = \tilde{\mathbf{S}}_{\boldsymbol{\lambda}} + \boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ or $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}} = \tilde{\mathbf{S}}_{\boldsymbol{\lambda}} + \boldsymbol{\Lambda}^{\mathrm{R}}_{\lambda_{\vartheta^*}}$ and $\bar{\boldsymbol{\lambda}}$ includes both $\boldsymbol{\lambda}$ and λ_{ϑ^*}. Problem (5.6) can now be solved using the approach described in Section 3 where matrix $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ is replaced by $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}$. If $\mathcal{P}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathbf{0}$ then $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}$ clearly reduces to $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$.

The asymptotic behavior of the proposed estimator is detailed in Supplementary Material K for the sake of space.

5.2 *Simulation Study II*

The aim of this simulation study is to assess the performance of the correlation-based penalty approach described above. We will use DGP2 from Section 4.1.2. Finally, the effectiveness of the method in estimating smooth function components will be explored.

5.2.1 *DGP2* Recall from Simulation Study I in Section 4.1 that the correlation parameter estimates were not deemed satisfactory at $n = 1000$. Here, we re-examine this case by employing trivariate probit models with penalized correlations, using

```
outR  <- SemiParTRIV(f.l, data = dat, penCor = "ridge" )
outL  <- SemiParTRIV(f.l, data = dat, penCor = "lasso" )
outAL <- SemiParTRIV(f.l, data = dat, w.alasso = w.al, penCor = "alasso")
```

where `f.l` and `data` are defined in Section 4.1. Argument `penCor` specifies the type of penalty used for the correlation parameters (`ridge`, `lasso` or `alasso`) and `w.alasso` denotes a 3×1 vector including the adaptive weights chosen as

```
w.al = c(theta12.ML, theta13.ML, theta23.ML)
```

where `theta12.ML, theta13.ML` and `theta23.ML` correspond to $\hat{\vartheta}_{12}^{\text{MLE}}$, $\hat{\vartheta}_{13}^{\text{MLE}}$ and $\hat{\vartheta}_{23}^{\text{MLE}}$. Table 2 shows substantial gains in accuracy and precision when penalizing the correlation parameters. In this case, using `lasso` produced better overall performances as compared to `alasso` and `ridge`, although such differences may be judged as negligible.

5.2.2 *DGP3* To assess the ability of `SemiParTRIV()` in estimating smooth function components, we modified slightly DGP2 by introducing non-linear effects for the continuous variable in the model. Estimation was achieved using the same syntax as that shown in the previous section but with equations specified as

```
eqn1 ~ v1 + s(z1); eqn2 ~ v1 + s(z1); eqn3 ~ v1 + s(z1)
```

where `s(z1)` defines a smooth function of the continuous covariate `z1`. A detailed description of DGP3 as well as the corresponding `R` code can be found in Supplementary Material I.2. In this case the coefficients of the spline bases and the correlations were penalized. The Lasso-type correlation-based penalty was employed (using Ridge and Adaptive Lasso produced virtually identical results). The estimates for

the correlations and parametric part of the model were very similar to those of the previous study. Figure 3 shows that the estimated curves recover the true functions reasonably well. For $n = 1000$, the estimates are rather variable and there are cases where the estimated functions are either wigglier or smoother than they should be. This does not come as a surprise recalling that we are dealing with simultaneous binary models and as the sample size grows large the results improve considerably. Finally, we calculated 95% average coverage probabilities for the model's smooth functions using point-wise intervals based on the result mentioned in Section 4. The coverages for $s_1(z_1)$, $s_2(z_1)$ and $s_3(z_1)$ were 0.959, 0.956 and 0.974 for $n = 1000$, and 0.949, 0.950 and 0.951 for $n = 10000$, hence confirming the good performance of the employed approximation.

The proposed approach generally proved effective. However, one should bear in mind that if the observed proportions of some trivariate binary events are very low then estimation may become challenging if not infeasible in some cases.

6. Analysis of North Carolina data

Birth weight and gestational age are strongly related with infant morbidity and mortality (Paneth, 1995; Butler et al., 2007). Infant's low birth weight (LBW) and preterm birth (PTB) are typically defined as binary variables taking value 1 when weight is less than 2500 grams, and number of gestation weeks is less than 37, respectively (e.g., Neelon et al., 2014). Kiely (1998) and Martin et al. (1999) argued that multiple birth (MB), also modeled as a binary variable, is strongly related with PTB and LBW. These variables are typically influenced by geographic, demographic and behavioral characteristics (e.g., Neelon et al., 2014; Miranda et al., 2009; South et al., 2012). This section illustrates the proposed modeling framework using 2007-2008 birth data from the North Carolina Center for Health Statistics (`http://www.schs.state.nc.us/`). The goal is to analyse jointly LBW, PTB and MB conditional on flexible functions of covariates and to account for residual dependence between the responses.

6.1 *Model specifications and results*

The data set consists of $61,426$ female newborns (similar results were obtained for male infants) which provides details on infant, maternal health and parental characteristics. The choice of variables included in the model was mainly driven by previous work on the subject (e.g., Miranda et al., 2009; South et al., 2012; Neelon et al., 2014). The responses are plurality (`mb`), a binary variable that takes value 1 for singleton birth and 0 otherwise, infant's birth weight (`lbw`) and preterm birth (`ptb`) which have been defined above. The covariates are maternal race categorized as non-white and white (`nwhite`), smoking status with 1 indicating a mother smoking during pregnancy (`smoker`), weight gained by mother during pregnancy in pounds (`gained`), age of mother in years (`mage`) and county in which the birth occurred (`county`). We employed `STATA`'s function `mvprobit()` and the proposed `SemiParTRIV()`. The model equations are

$$\begin{aligned}
\texttt{mb}_i^* &= \beta_{11} + \beta_{12}\texttt{nwhite}_i + \beta_{13}\texttt{smoker}_i + \texttt{gained}_i + \texttt{mage}_i + \texttt{county}_i + \varepsilon_{1i},\\
\texttt{lbw}_i^* &= \beta_{21} + \beta_{22}\texttt{nwhite}_i + \beta_{23}\texttt{smoker}_i + \texttt{gained}_i + \texttt{mage}_i + \texttt{county}_i + \varepsilon_{2i},\\
\texttt{ptb}_i^* &= \beta_{31} + \beta_{32}\texttt{nwhite}_i + \beta_{33}\texttt{smoker}_i + \texttt{gained}_i + \texttt{mage}_i + \texttt{county}_i + \varepsilon_{3i}.
\end{aligned}$$

The regression coefficient estimates for the two competing methods were very similar and are not reported here. However, as shown in Table 3, the estimated correlations are significantly different. Moreover, the proposed approach was faster and produced narrower intervals as compared to those of `STATA`'s routine. Figure 4 depicts the joint probabilities (averaged by county) that birth is multiple, infant's birth weight is normal and the baby is born full term. The probabilities obtained using `mvprobit()` are overall higher than those obtained using `SemiParTRIV()`. This can be attributed to the different correlation estimates of the two methods. Our simulations showed that `STATA`'s routine produces biased correlation estimates, hence we would be reluctant to trust such results.

Our approach allows for flexible functional dependence of the responses on the covariates. We there-

fore re-specify the model using the following equations

$$\texttt{mb}_i^* = \beta_{11} + \beta_{12}\texttt{nwhite}_i + \beta_{13}\texttt{smoker}_i + s_{11}(\texttt{gained}_i) + s_{12}(\texttt{mage}_i) + s_{1\texttt{spatial}}(\texttt{county}_i) + \varepsilon_{1i},$$

$$\texttt{lbw}_i^* = \beta_{21} + \beta_{22}\texttt{nwhite}_i + \beta_{23}\texttt{smoker}_i + s_{21}(\texttt{gained}_i) + s_{22}(\texttt{mage}_i) + s_{2\texttt{spatial}}(\texttt{county}_i) + \varepsilon_{2i},$$

$$\texttt{ptb}_i^* = \beta_{31} + \beta_{32}\texttt{nwhite}_i + \beta_{33}\texttt{smoker}_i + s_{31}(\texttt{gained}_i) + s_{32}(\texttt{mage}_i) + s_{3\texttt{spatial}}(\texttt{county}_i) + \varepsilon_{3i},$$

where s_{m1} and s_{m2}, $\forall m = 1, 2, 3$, are smooth functions of $\texttt{gained}_i$ and $\texttt{mage}_i$ represented using penalized thin plate regression splines with 20 base functions and second order penalties, and $s_{m\texttt{spatial}}$ models spatial regional effects using a Markov random field approach.

An example of estimated regression effects is shown in Figure 5 for the `lbw` equation. This suggests that the likelihood of low birth weight decreases with weight gained by the mother during pregnancy (with a pick at around 40 pounds) and then increases (although with quite some uncertainty). The effect of mother's age on the propensity of lower infant's birth weight appears to be almost steady up to 30 years with a dramatic increase for women older than 40 years. Note that the estimated smooths are centered around zero because of centering identifiability constraints (see Section 2), however this does not affect interpretation. The point-wise confidence intervals do not contain the zero line in most of the ranges of the `gained` and `mage` values. This suggests that these two variables are important factors in determining `lbw`. The spatial map shows the effects of the county variable on the outcome, where darker colors correspond to a decreased propensity of low birth weight. P-values for testing smooth components for equality to zero were obtained by adapting the results discussed in Wood (2013a) and Wood (2013b) to the current context. These showed that the covariate effects are significant at least at the 5% level.

7. Discussion

We have introduced a penalized likelihood method to estimate a trivariate system of probit regressions that incorporate additive or semi-parametric effects. The approach can also penalize the model's correlation coefficients via differentiable and approximations of non-differentiable penalties. This addresses the

difficulty in estimating accurately the correlation parameters at small or modest sample sizes, an issue that has been neglected in the literature and that is likely to have a detrimental impact on the empirical performance of simultaneous binary models with more than two responses. The proposed developments are backed by a reliable estimation method which requires analytical information on the score vector and Hessian matrix of the model's log-likelihood. Such information is not readily available in the literature and has been provided in this paper. Some asymptotic properties of the proposed estimator have also been discussed. The proposed model can be easily fitted using the `SemiParTRIV()` function in the `R` package `SemiParBIVProbit`. The proposed method has been illustrated through simulations as well as a case study whose aim was to estimate a simultaneous model for three binary outcomes of newborn infants in North Carolina. Our results showed that joint outcome probabilities are affected by the way the model's parameters are estimated, especially the correlation coefficients.

Future work will look into the feasibility of modeling the correlation parameters as functions of flexible predictors, and into extending the material in Section 4 to accommodate link functions other than probit. Another interesting extension would be to exploit pair-copula and composite likelihood constructions to allow for non-Gaussian dependencies between the responses. A future release of `SemiParBIVProbit` will also incorporate the option of fitting trivariate probit models with double sample selection (e.g., Zhang et al., 2015); this will require deriving the model's log-likelihood and its respective score and Hessian components, but the proposed framework will be essentially unaffected by such changes.

REFERENCES

Ashford, J. & Sowden, R. (1970). Multi-variate probit analysis. *Biometrics*, 26, 535–546.

Azzalini, M. A. (2014). *The Multivariate Normal and t Distributions*. R package version 1.5-3.

Butler, A. S., Behrman, R. E., et al. (2007). *Preterm Birth: Causes, Consequences, and Prevention.* National Academies Press.

Cappellari, L. & Jenkins, S. P. (2003). Multivariate probit regression using simulated maximum likelihood. *The Stata Journal*, 3(3), 278–294.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2), 347–361.

Donat, F. & Marra, G. (2016). Semi-parametric bivariate polychotomous ordinal regression. *Statistics and Computing*.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2), 89–102.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

Genest, C., Nikoloulopoulos, A. K., Rivest, L.-P., & Fortin, M. (2013). Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. *Brazilian Journal of Probability and Statistics*, 27(3), 265–284.

Gilbert, P. & Varadhan, R. (2015). numderiv: Accurate numerical derivatives. *R package version 2014.2-1*.

Greene, W. (2003). *Econometric Analysis*. Prentice Hall, New York.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag, London.

Henningsen, A. (2015). *mvProbit: Multivariate Probit Models*. R package version 0.1-8.

Kiely, J. L. (1998). What is the population-based risk of preterm birth among twins and other multiples? *Clinical Obstetrics and Gynecology*, 41(1), 3–11.

Klein, N. & Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: A unifying bayesian approach. 26(4), 841–860.

Koch, I. (1996). On the asymptotic performance of median smoothers in image analysis and nonparametric regression. *The Annals of Statistics*, 24(4), 1648–1666.

Król, A., Ferrer, L., Pignon, J.-P., Proust-Lima, C., Ducreux, M., Bouché, O., Michiels, S., & Rondeau, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the ffcd 2000–05 trial. *Biometrics*.

LP, S. C. (2015). *Stata Statistical Software Release 13*. Stata Press Publication.

Marra, G. & Radice, R. (2017). *SemiParBIVProbit: Semiparametric Copula Regression Models*. R package version 3.8-1.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., McGovern, M. E., et al. (2016). A simultaneous equation approach to estimating hiv prevalence with non-ignorable missing responses. *Journal of the American Statistical Association*.

Martin, J. A., Park, M. M., et al. (1999). Trends in twin and triplet births: 1980–97. *National vital statistics reports*, 47(24), 1–16.

Miranda, M. L., Maxson, P., & Edwards, S. (2009). Environmental contributions to disparities in pregnancy outcomes. *Epidemiologic Reviews*, 31(1), 67–83.

Neelon, B., Anthopolos, R., & Miranda, M. L. (2014). A spatial bivariate probit model for correlated binary data with application to adverse birth outcomes. *Statistical Methods in Medical Research*, 23(2), 119–133.

Oelker, M.-R. & Tutz, G. (2013). *Technical Report Number 139, 2013, Department of Statistics, University of Munich*.

Paneth, N. S. (1995). The problem of low birth weight. *The Future of Children*, 5(1), 19–34.

Radice, R., Marra, G., & Wojtyś, M. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5), 981–995.

Rippe, R. C., Meulman, J. J., & Eilers, P. H. (2012). Visualization of genomic changes by segmented smoothing using an l_0 penalty. 7(6), 1–14.

Rousseeuw, P. J. & Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, 22(4), 965–984.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge university press.

South, A. P., Jones, D. E., Hall, E. S., Huo, S., Meinzen-Derr, J., Liu, L., & Greenberg, J. M. (2012). Spatial analysis of preterm birth demonstrates opportunities for targeted intervention. *Maternal and Child Health Journal*, 16(2), 470–478.

Team, R. D. C. (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Terracol, A. (2002). Triprobit and the ghk simulator: a short note.

Trinh, G. & Genz, A. (2015). Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing*, 25(5), 989–996.

Ulbricht, J. (2010). *Variable selection in generalized linear models*. Verlag Dr. Hut.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, London.

Wood, S. N. (2013a). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S. N. (2013b). A simple test for random effects in regression models. *Biometrika*, 100(4), 1005–1010.

Yoshida, T. & Naito, K. (2014). Asymptotics for penalised splines in generalised additive models. *Journal of Nonparametric Statistics*, 26(2), 269–289.

Zhang, R., Inder, B. A., & Zhang, X. (2015). Bayesian estimation of a discrete response model with double rules of sample selection. *Computational Statistics & Data Analysis*, 86, 81–96.

Zhong, W., Koopmeiners, J. S., & Carlin, B. P. (2012). A trivariate continual reassessment method for phase i/ii trials of toxicity, efficacy, and surrogate efficacy. *Statistics in Medicine*, 31(29), 3885–3895.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

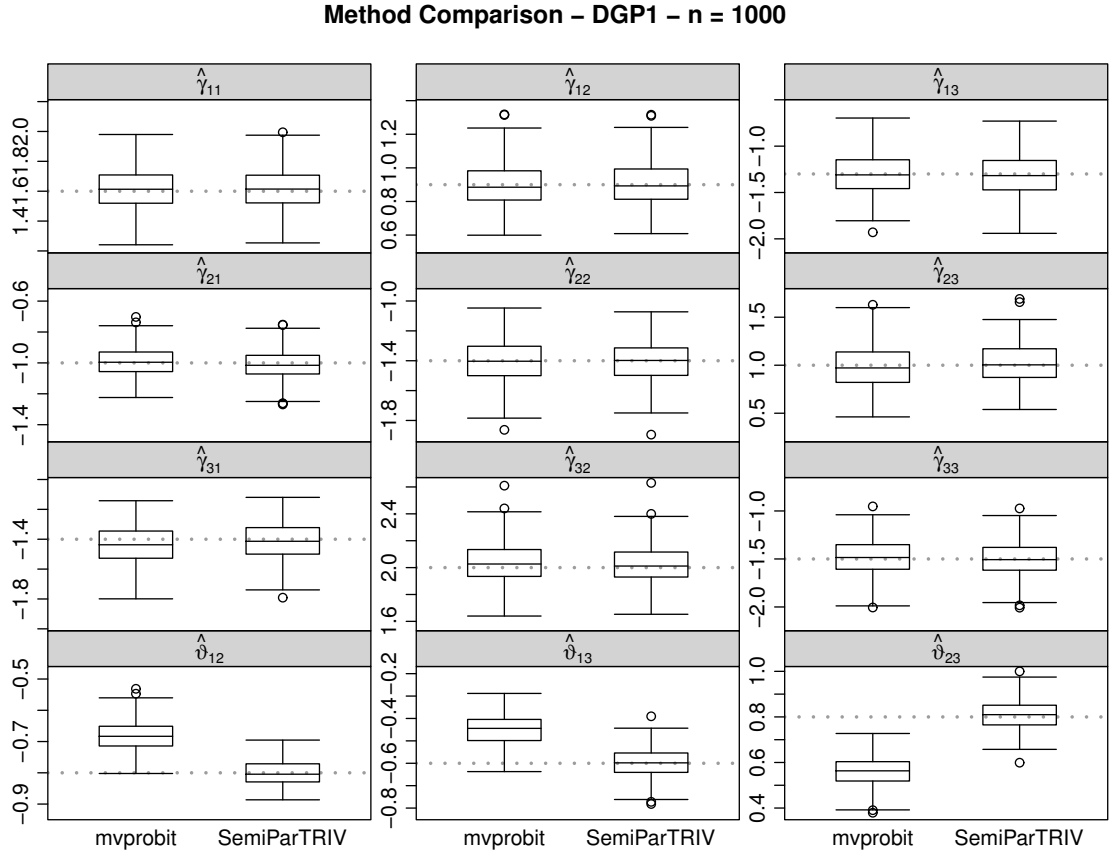

Fig. 1: Boxplots of parameter estimates obtained by applying `mvprobit()` and `SemiParTRIV()` to 250 datasets simulated according to DGP1. The sample size was equal to 1000 and the true parameter values are represented by horizontal gray dotted lines.

	DGP2			
	$n = 1000$		$n = 10000$	
Estimator	Bias (%)	RMSE	Bias (%)	RMSE
$\hat{\vartheta}_{12}$	11.36	0.0935	-0.79	0.0262
$\hat{\vartheta}_{13}$	13.53	0.1204	1.86	0.0320
$\hat{\vartheta}_{23}$	-2.02	0.0567	0.16	0.0129

Table 1: Percentage biases and root mean squared errors of the correlation estimates obtained by applying `SemiParTRIV()` to 250 datasets simulated according to DGP2. $\text{RMSE}(\hat{\vartheta}_{zk})$ is given by $\sqrt{\frac{1}{250}\sum_{\iota=1}^{250}\{\hat{\vartheta}_{zk,\iota} - \vartheta_{zk0}\}^2}$ where $\hat{\vartheta}_{zk,\iota}$ denotes the ι-th estimated value and ϑ_{zk0} is the true one.

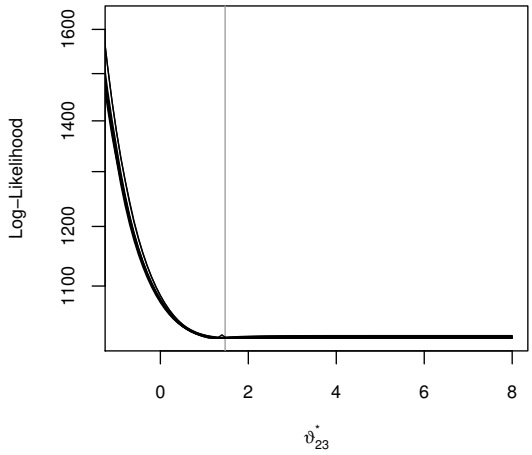

Fig. 2: Profile log-likelihood function of the trivariate probit model for correlation parameter ϑ^*_{23}, for 10 data sets of sample size 1000 generated using DGP2. The true value is represented by the vertical grey line.

		DGP2, $n = 1000$	
Estimator	Correlation-based penalty	Bias (%)	RMSE
$\hat{\vartheta}_{12}$	Unpenalized	11.36	0.0935
	Ridge	0.10	0.0903
	Lasso	0.02	0.0835
	Adaptive Lasso	-0.31	0.0862
$\hat{\vartheta}_{13}$	Unpenalized	13.53	0.1204
	Ridge	0.13	0.1158
	Lasso	0.07	0.1092
	Adaptive Lasso	0.03	0.1142
$\hat{\vartheta}_{23}$	Unpenalized	-2.02	0.0567
	Ridge	-0.03	0.0551
	Lasso	-0.02	0.0475
	Adaptive Lasso	0.01	0.0428

Table 2: Percentage biases and root mean squared errors of the correlation estimates obtained by applying `SemiParTRIV()` to 250 datasets simulated according to DGP2 when the unpenalized approach as well as Ridge, Lasso and Adaptive Lasso correlation-based penalties are employed.

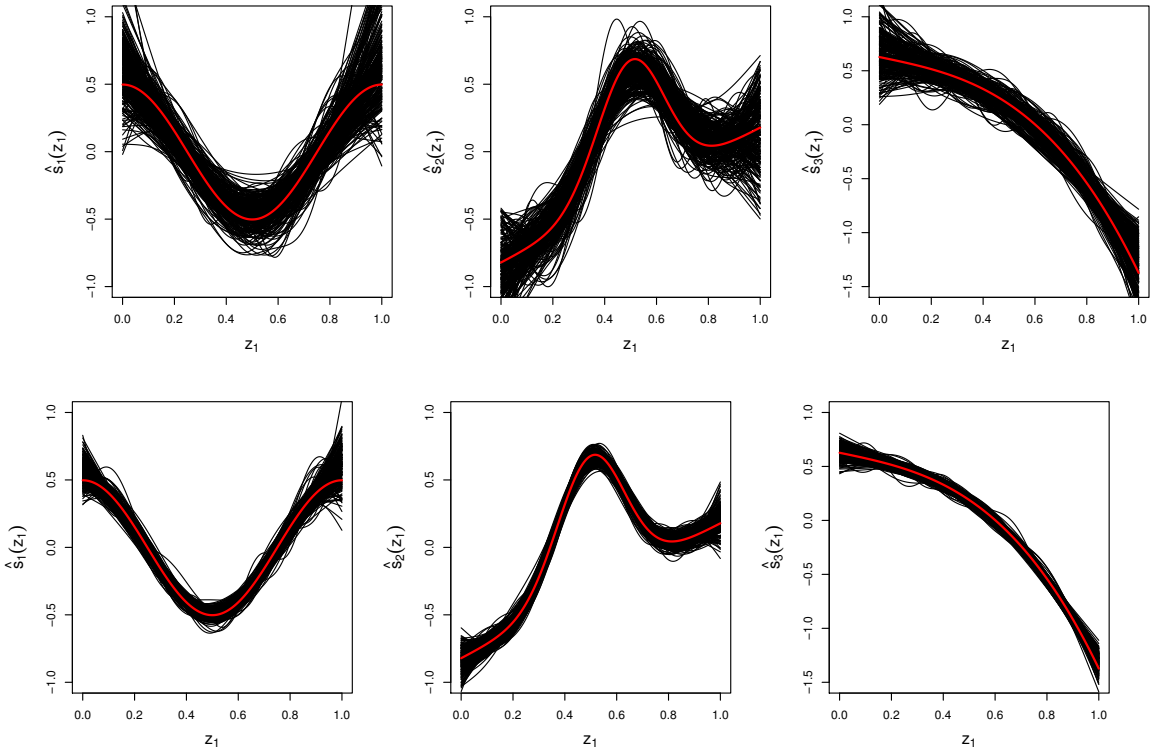

Fig. 3: Estimated smooth functions for $s_1(z_1)$, $s_2(z_1)$ and $s_3(z_1)$ obtained by applying `SemiParTRIV()` to 250 datasets simulated according to DGP3. The first row shows the estimated curves obtained from samples of 1000 observations, whereas those in the second row correspond to samples of 10000 observations. The black lines represent the estimated smooth functions over all replicates and the red solid lines refer to the true functions.

	`SemiParTRIV()`	`mvprobit()`
$\hat{\vartheta}_{12}$ (95% CI)	-0.7617 $(-0.7612, -0.7622)$	-0.5191 $(-0.5027, -0.5351)$
$\hat{\vartheta}_{13}$ (95% CI)	-0.6397 $(-0.6390, -0.6402)$	-0.4277 $(-0.4107, -0.4443)$
$\hat{\vartheta}_{23}$ (95% CI)	0.7853 $(0.7850, 0.7856)$	0.6796 $(0.6692, 0.6897)$
Execution Time	296.26	349.41

Table 3: Correlation parameter estimates obtained by using `SemiParTRIV()` and `mvprobit()`. Corresponding 95% intervals (CIs) are reported in parentheses. The execution time (in seconds) for each method is reported at the bottom of the table.

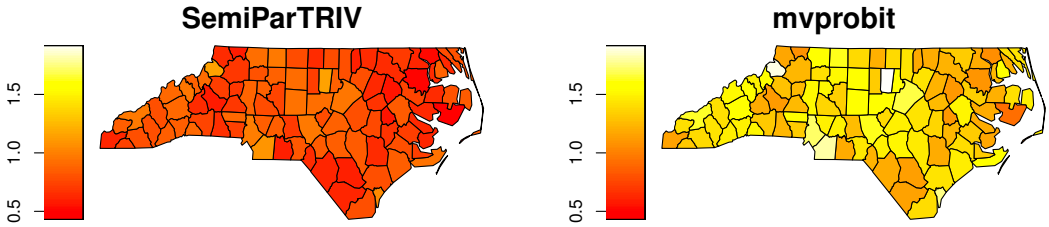

Fig. 4: Joint probabilities (in %) that `mb` is multiple, `lbw` is > 2500 grams and `ptb` is > 37 weeks by county in North Carolina, obtained using by `SemiParTRIV()` and `mvprobit()`.

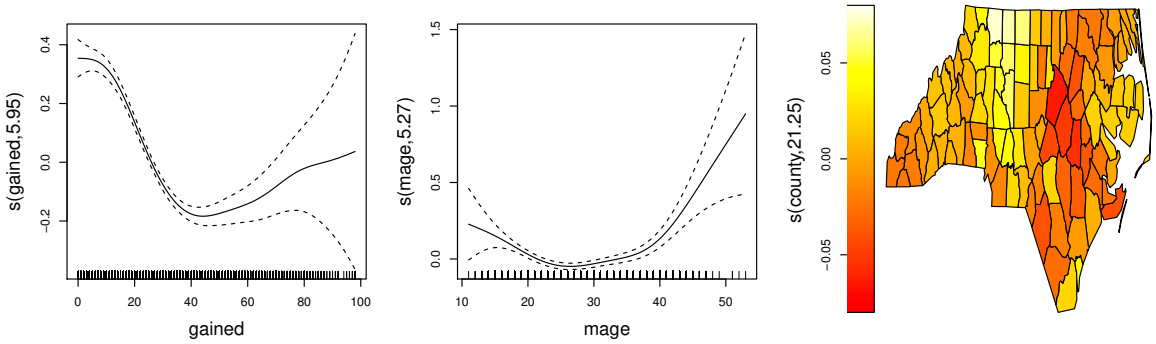

Fig. 5: Smooth effects of `gained` and `mage` on `lbw` and associated 95% point-wise confidence intervals. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in brackets in the y-axis captions specify the *edf* of the smooth curve ($edf = 1$ corresponds to a straight line estimate; the higher the value the more complex the estimated curve). The map on the right hand side shows the magnitude of the regional variable in each of the 100 counties in North Carolina.

[*Received XX; revised XX; ...*]